Bootstrap tests for misspecified models, 
with application to clustered binary data

Marc Aerts and Gerda Claeskens
Center for Statistics, Limburgs Universitair Centrum,
Universitaire Campus, B-3590 Diepenbeek, Belgium

Summary

When the data do not come from the assumed parametric model, the usual asymptotic chi-
squared distribution under the null hypothesis, remains valid for “robustified” Wald and score 
test statistics. In this paper we compare the performance of this chi-squared approximation to
that of a semiparametric bootstrap method. The bootstrap approximation is based on a one-
step bootstrap estimator reflecting the null hypothesis. One of the advantages of this one-step
approach is that no bootstrap data have to be generated and no additional model fitting is
required. Simulations on clustered binary data indicate that the robust score test is superior
and that, in cases where the chi-squared type tests fail in reaching the prescribed significance
level, the proposed bootstrap test succeeds in correcting this towards the nominal level. The
different methods are also compared on real developmental toxicity data.

Key words: Clustered binary data; Developmental toxicity; Hypothesis testing; Model misspec-
ification, Pseudolikelihood, Semiparametric bootstrap.
1 Introduction

This work is mainly motivated by toxicological experiments which are designed to assess the potential adverse effects of drugs or other exposures on developing fetuses of pregnant rodents (usually mice or rats). A typical study includes a control group and some dosed groups. The exposure occurs early in gestation, the animals are sacrificed prior to term and the uterine contents are examined for malformations. Since littermates are likely to show the same behaviour, this kind of experiments results in cluster correlated binary outcomes (malformation: yes or no). Emphasis can be placed on estimating a dose effect parameter, on testing the null hypothesis of no dose effect, or on determining a benchmark dose. The example discussed in this paper concerns a study conducted by the Research Triangle Institute under contract to the National Toxicology Program (NTP). It investigates the effect in rats and mice of the chemical theophylline (Lindstrom et.al., 1990). The main question here is whether exposure to theophylline has an adverse effect on developing fetuses, so focus is on testing the null hypothesis of no effect.

For clustered binary data from experiments like this one, it is often not clear which probability model is generating the particular data. Different types of models (marginal, conditional, random effects models) are available and estimation methods range from full likelihood to pseudolikelihood, quasilikelihood and generalized estimating equations. A thorough review is given in Pendergast et al. (1996). In this paper we are interested in methods based on the assumption that a specific probability model holds: full likelihood and pseudolikelihood as in Besag (1975), Cressie (1991) and for clustered binary data in Geys, Molenberghs and Ryan (1997, 1999). If the probability model is misspecified, it is well-known that classical test statistics (such as the Wald, score or likelihood ratio statistic) do not have an asymptotically chi-squared distribution anymore (see, e.g., White 1982). By means of asymptotic calculations, small sample simulations and analyses of NTP data, Molenberghs, Declerck and Aerts (1998) investigated the behaviour of the parameter estimates and the classical Wald and likelihood ratio test under model misspecification, while still using the incorrect asymptotic chi-squared distribution. It is desirable, however, to use modified test statistics, for robustifying the standard inference methods. For full likelihood models, robust Wald and score tests have been described by, e.g., Kent (1982), Viraswami and Reid (1996). The modified tests again have an asymptotic chi-squared distribution, even when the assumed model is not correct. We are not aware of a modified likelihood ratio test with asymptotic chi-squared distribution. Robust test statistics are also used in the context of generalized estimating equations (only specifying the mean and variance structure, see Liang and Zeger 1986, Rotnitzky and Jewell 1990) and in the pseudolikelihood approach, see Geys, Molenberghs and Ryan (1999).

In this paper we focus attention on the likelihood and pseudolikelihood approach and we
investigate, when using a misspecified probability model, the performance of the robust Wald and score test statistic based on the chi-squared approximation and based on a bootstrap estimator for their distribution. The methodology and theory for bootstrap hypothesis testing is still not fully developed. The main difficulty is the generation of bootstrap data reflecting the null hypothesis. Assuming the true likelihood of the data to be known, this can be achieved by the parametric bootstrap based on the null estimates. For clustered binary data models, Aerts and Claeskens (1999) have shown that the parametric bootstrap test leads to a substantial improvement. In practice however the assumed probability model can be wrong, in which case the parametric bootstrap leads to incorrect results. For complex data structures, there is (at least by our knowledge) no general nonparametric or semiparametric method available and “theoretical and empirical studies are still called for” (citing Shao and Tu, 1995, p.189).

As a possible approach, we propose a semiparametric bootstrap method. It remains valid when the assumed model is incorrect and no bootstrap data are generated such that no iterative fitting is required. Instead bootstrap parameter estimates reflecting the null hypothesis are generated directly. Our approach is closely related to the one-step bootstrap, an approximate method to simplify the bootstrap for estimators which have to be computed iteratively (see Schucany and Wang 1991 and Section 5.4.7 in Shao and Tu 1995). The method is described in Section 4, including a second order improvement. The results of a simulation study and a data example are shown in Section 5. In particular, the developmental toxicity data on theophylline are analyzed on possible dose effect. It should be noted that although we restrict attention to (pseudo)likelihood methods for clustered binary data, the technique can be applied to generalized estimating equations (GEE). For the GEE estimation method however, no probability distribution has to be specified and hence, by nature of the method, it is expected to behave more robust against misspecification. For illustration we also included results of the GEE method in the analysis of the NTP data.

In the next section, we briefly introduce the pseudolikelihood method together with some notation and in Section 3 those clustered binary data models which are used for the simulations and data analyses are briefly reviewed.

2 Pseudolikelihood and misspecification

Let $Y_{i1}, \ldots, Y_{im_i}$ be independent identically distributed random variables of length $m$ with common (unknown) joint density or discrete probability function (pdf) $g_i(y), y = (y_1, \ldots, y_m) \in R_i, i = 1, \ldots, p$. The number $p$ of possibly different (associated) populations is considered as fixed whereas the number $n_i$ of observations from the distinct populations become large as $n = \sum_{i=1}^{p} n_i$ tends to infinity, according to $n_i/n \to \lambda_i$ where $\sum_{i=1}^{p} \lambda_i = 1$ with $\lambda_i > 0$. 

In the context of clustered binary data from toxicological experiments, the different populations correspond to different dose levels \( d_i \), \( n_i \) is the number of litters exposed to dose \( d_i \), \( m \) is the litter size and \( y_{ijk} \) indicates whether the \( k \)th fetus of litter \( j \) in population \( i \) is malformed or not. For ease of notation, we restrict to a fixed littersize \( m \). In real situations, the size \( m \) varies among the litters.

In general, parametric inference for associated populations is based on \( r \) dimensional vector functions \( \psi_i(y, t) \), the score functions, where the “true” parameter \( \theta = (\theta_1, \ldots, \theta_r) \) is defined as the solution \( t \) to

\[
\sum_{i=1}^{p} \lambda_i E[\psi_i(Y_{i1}; t)] = 0, \tag{1}
\]

where all expectations are w.r.t. the true pdf \( g_i(y) \).

Solving the system of equations

\[
\sum_{i=1}^{p} \sum_{j=1}^{n_i} \psi_i(Y_{ij}; t) = 0, \tag{2}
\]

leads to the estimator \( \hat{\theta}_n \) for \( \theta \).

We focus attention on pseudolikelihood methods. These are based on a fully specified joint pdf \( f_i(y, \theta) \), possibly different from \( g_i(y) \). Define \( S_i \) as the set of all \( 2^m - 1 \) vectors \( s \) of length \( m \), consisting solely of zeros and ones, with each vector having at least one nonzero entry. Denote by \( Y_{ij}^{(s)} \) the subvector of \( Y_{ij} \) corresponding to the non-zero components of \( s \) with associated joint density function \( f_i^{(s)}(y^{(s)}/\theta) \). The log of the pseudolikelihood is defined as

\[
\log PL_n(\theta) = \sum_{i=1}^{p} \sum_{s \in S_i} \delta_s \sum_{j=1}^{n_i} \log f_i^{(s)}(y^{(s)}/\theta)
\]

with \( \{\delta_s | s \in S_i\} \) a set of \( 2^m - 1 \) real numbers, not all zero. Classical maximum likelihood corresponds to \( \delta_s = 1 \) for \( s = 1_m \) and zero otherwise, where \( 1_m \) is a vector of ones. Another typical choice is \( \delta_{1_m} = m \) and \( \delta_{s_\ell} = -1 \) for \( \ell = 1, \ldots, m \) where \( s_\ell \) consists of ones everywhere, except for the \( \ell \)th entry. This particular choice is referred to as the “full conditional” log pseudolikelihood function. It has the effect of replacing the joint density function by a product of \( m \) univariate conditional density functions, thus avoiding the incorporation of a possibly complicated normalizing constant which typically arises in exponential family models (an example is given in Section 3).

For the pseudolikelihood method, the score functions \( \psi_i \) are the partial derivatives

\[
\psi_i(y, t) = \sum_{s \in S_i} \delta_s \frac{\partial}{\partial t} \log f_i^{(s)}(y^{(s)}, t).
\]

There can be different sources of misspecification: an incorrect pdf and/or pseudolikelihood instead of full likelihood. Within the classical maximum likelihood approach where \( \psi_i(y, t) = \)
(\partial/\partial t) \log f_i(y, t), the assumed pdf \( f_i(y, t) \) might not contain the true structure \( g_i(y) \). In this case, assuming interchangeability of integration and derivation, the parameter \( \theta \) defined by (1) is that value in the parameter space \( \Theta \subset \mathbb{R}^r \) which brings \( f_i(y, t) \) as close as possible to \( g_i(y) \); that is, \( \theta \) minimizes the Kullback-Leibler information criterion:

\[ \sum_{i=1}^p \lambda_i E \left[ \log \left( \frac{g_i(Y_{i1})}{f_i(Y_{i1}, t)} \right) \right]. \]

Under severe model misspecification, this might lead to parameters \( \theta \) without direct biologically meaningful interpretation. Therefore, great care should be taken in the formulation of the null hypothesis to be tested.

### 3 Models for clustered binary data

Different types of models are available for correlated binary data. We restrict attention to models for univariate clustered outcomes. As examples of full likelihood models, we selected the beta-binomial model, which can be viewed as a random effects model, and the conditional model (MR) of Molenberghs and Ryan (1999). Next to these models, we also examined the pseudolikelihood model of Geys, Molenberghs and Ryan (1997).

#### 3.1 A conditional likelihood model

Molenberghs and Ryan (1999) proposed a likelihood model for multiple clustered binary outcomes, with following probability density function (restricted to the special case of a univariate clustered outcome):

\[ f_i(y, \theta) = \exp \{ (z, -z(m - z)) X_i \theta - A(\theta) \}, \]

where \( y = (y_1, \ldots, y_m) \) \( (y_j \) indicates whether the \( j \)th individual in the cluster is malformed\), \( z \) is the total number of malformations in the cluster and \( A(\theta) \) is a normalizing constant. \( X_i \) is the design matrix based on dose \( d_i \) associated with this cluster. For example,

\[ X_i = \begin{pmatrix} 1 & d_i & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad \theta = \begin{pmatrix} \theta_{10} \\ \theta_{11} \\ \theta_{20} \end{pmatrix} \]

which corresponds to a linear effect on the main parameter (the coefficient of \( z \)) and a constant intra-litter association (the coefficient of \(-z(m - z)\)). This model is based on the multivariate exponential family model proposed by Cox (1972). It benefits from the elegance and simplicity of exponential family theory. More details about model properties and inference can be found in Molenberghs and Ryan (1999).
3.2 A pseudolikelihood model

As an exponential family model, the Molenberghs-Ryan conditional model enjoys well-known properties, including numerical stability. This computational advantage can be lost, especially with larger clusters where the calculation of the normalizing constant $A(\theta)$ can be quite cumbersome. To overcome this problem, Geys et al. (1997, 1999) propose the use of pseudolikelihood. The joint probability density function $f_i(y, \theta)$ is now being replaced by

$$
\prod_{k=1}^{m} f_i(y_k | y_{\ell} \neq k, \theta) = \left[ f_i(y, \theta) \right]^{m} \prod_{k=1}^{m} f^{(s_k)}_i(y^{(s_k)}, \theta), \tag{5}
$$

Equation (5) represents the product of all $m$ conditional probabilities of the outcome of one fetus, given all other observations. This is the full conditional pseudolikelihood as mentioned in Section 2. An equivalent way of writing this model is as follows:

$$
\prod_{k=1}^{m} f_i(y_k | y_{\ell} \neq k; \xi, \psi) = \prod_{k=1}^{m} p_{is}^{y_k} p_{if}^{1-y_k}
$$

where $p_{is}$ is the conditional probability of an additional success,

$$
\text{logit}(p_{is}) = \text{logit}\{ P(y_k = 1 | z - 1 \text{ successes} \& m - z \text{ failures}) \} = (1 - (m - 2z + 1))X_i\theta
$$

and $p_{if}$ is the conditional probability of an additional failure,

$$
\text{logit}(p_{if}) = \text{logit}\{ P(y_k = -1 | z \text{ successes} \& m - z - 1 \text{ failures}) \} = (-1 - (m - 2z - 1))X_i\theta.
$$

The contribution of this cluster to the log pseudolikelihood is then given by $z \log(p_{is}) + (m - z) \log(p_{if})$. More information on the pseudolikelihood model can be found in Geys, Molenberghs and Ryan (1997, 1999).

3.3 A random effects likelihood model

Next to the MR-model and its pseudolikelihood version, we used the well-established beta-binomial model (Skellam 1948). The beta-binomial approach assumes a random malformation probability $P$ in a cluster to come from a beta distribution with mean $\pi$. Given $P$, the outcomes within a cluster follow a binomial distribution (see, e.g, Kleinman 1973). Its probability density function is given by

$$
f_i(y, (\pi, \rho)) = \binom{m}{z} \frac{B(\pi \rho^{-1} - 1 + z, (1 - \pi)(\rho^{-1} - 1) + (m - z))}{B(\pi \rho^{-1} - 1, (1 - \pi)(\rho^{-1} - 1))} \tag{6}
$$

where $\rho$ is the intraclass correlation (assuming exchangeability) and $B(., .)$ denotes the beta function. It can be shown that the log-likelihood is given by

$$
\ln \left( \binom{m}{z} \right) + \sum_{r=0}^{z-1} \ln \left( \pi + \frac{r \rho}{1 - \rho} \right) + \sum_{r=0}^{m-z-1} \ln \left( 1 - \pi + \frac{r \rho}{1 - \rho} \right) - \sum_{r=0}^{m-1} \ln \left( 1 + \frac{r \rho}{1 - \rho} \right).
$$
Note that this expression reduces to the familiar binomial log-likelihood when \( \rho = 0 \). The marginal parameters \( \pi \) and \( \rho \) are modeled as a function of dose \( d_i \) via the logistic link function for \( \pi \) and Fisher’s \( z \)-transform for \( \rho \). As an example, consider a linear dose effect on logit(\( \pi \)) and a constant intralitter correlation

\[
\begin{pmatrix}
\ln\left( \frac{\pi}{1-\pi} \right) \\
\ln\left( \frac{1+\rho}{1-\rho} \right)
\end{pmatrix}
= X_i \theta
\]

with \( X_i \) and \( \theta \) as in (4). Note that in case of no intralitter association (\( \theta_{20} = 0 \)), all three models (3), (5) and (6) coincide and reduce to ordinary logistic regression.

4 Testing hypotheses

Consider the hypothesis \( H_0 : \theta \in \Theta_0 \) versus \( H_1 : \theta \in \Theta \setminus \Theta_0 \) where \( \Theta_0 \) is a \( (r - t) \) dimensional subspace of the parameter space \( \Theta \) such that the parameter of interest \( \theta = (\theta_1, \ldots, \theta_r) \) belongs to \( \Theta_0 \) if and only if \( \theta_1 = \ldots = \theta_t = 0, 1 \leq t \leq r \). More general situations, in which \( H_0 \) is of the form \( H_0 : h_1(\theta) = \ldots = h_t(\theta) = 0 \) for some smooth real-valued functions \( h_1, \ldots, h_t \), can be put into this form by a reparametrisation. Note that the hypothesis is stated in terms of the parameter which minimizes the Kullback-Leibler information criterion. The proposed method is presented in full generality but, in general, when misspecification might occur, one always has to decide whether the stated hypothesis is meaningful.

4.1 Robustified test statistics

If model misspecification arises, classical Wald, score and likelihood ratio tests do not have an asymptotic chi-squared distribution anymore. This is due to the fact that the Bartlett identities are no longer valid, e.g.,

\[
B_i(\theta) \equiv E \left[ \psi_i(y, \theta) \psi_i(y, \theta)^T \right] \neq -E \left[ \frac{\partial}{\partial \theta} \psi_i(y, \theta) \right] \equiv A_i(\theta).
\]

Wald and score tests can be robustified by using the so-called sandwich variance estimator. Properties of these robust test statistics are studied by Huber (1967), White (1982), Rotnitzky and Jewell (1990), Boos (1992) and Viraswami and Reid (1996) among others. There is no robustified version of the likelihood ratio test, whose asymptotic distribution is, under misspecification, a weighted sum of independent chi-squared random variables with one degree of freedom, where the weights are unknown and have to be estimated from the data. Properties of this test are studied by, e.g., Foutz and Srivastava (1977) and Kent (1982). We will not consider the likelihood ratio test.

Let us now introduce some notation. For a \( r \times 1 \) vector \( v \), let \( v_L \) denote the subvector of the first \( t \) components and for a \( r \times r \) matrix \( V \), define \( V_{LL} = C^T V C \) where \( C^T = [I_t, 0_{t,r-t}] \).
with $I_t$ the $t \times t$ identity matrix and $0_{t,r-t}$ the zero matrix of dimension $t \times (r-t)$. Denote by $V^{-1}_{LL}$ the inverse of the submatrix $V_{LL}$. The following matrices $A_n(\theta)$ and $B_n(\theta)$ will be used in the construction of the covariance matrix of the estimator $\hat{\theta}_n$,

$$A_n(\theta) = -n^{-1} \sum_{i=1}^{p} \sum_{j=1}^{n_i} \frac{\partial}{\partial \theta} \psi_i(Y_{ij}, \theta),$$

$$B_n(\theta) = n^{-1} \sum_{i=1}^{p} \sum_{j=1}^{n_i} \psi_i(Y_{ij}, \theta) \psi_i(Y_{ij}, \theta)^T.$$ 

The robust Wald and score test statistics

$$W_n = n(\hat{\theta}_n)_L^T \left( A_n(\hat{\theta}_n)^{-1} B_n(\hat{\theta}_n) A_n(\hat{\theta}_n)^{-1} \right)_{LL}^{-1} (\hat{\theta}_n)_L$$

$$S_n = \frac{1}{n} \left( \sum_{i=1}^{p} \sum_{j=1}^{n_i} \psi_i(Y_{ij}; \hat{\theta}_n^{(0)})_L \right)^T \left( A_n(\hat{\theta}_n^{(0)})^{-1} \right)_{LL} \left( A_n(\hat{\theta}_n^{(0)})^{-1} B_n(\hat{\theta}_n^{(0)}) A_n(\hat{\theta}_n^{(0)})^{-1} \right)_{LL}^{-1}$$

$$\times \left( A_n(\hat{\theta}_n^{(0)})^{-1} \right)_{LL} \left( \sum_{i=1}^{p} \sum_{j=1}^{n_i} \psi_i(Y_{ij}; \hat{\theta}_n^{(0)})_L \right)$$

each have an asymptotic $\chi^2_t$ distribution (White 1982).

The distributional properties of the robust Wald and score test rely on asymptotic approximations. Our goal is to examine how well these approximations work in the context of misspecified clustered binary data models and to study the performance of a semiparametric bootstrap test.

### 4.2 Bootstrap test statistics

The main difficulty when constructing bootstrap tests is the generation of bootstrap data reflecting the null hypothesis. Assuming that $g_i(y) = f_i(y, \theta)$, bootstrap data $\{Y^{*}_{ij}\}$ can be generated from the fitted model $f_i(y, \hat{\theta}_n^{(0)})$ where $\hat{\theta}_n^{(0)}$ is a $\sqrt{n}$-consistent estimator of $\theta$ under the null hypothesis. This parametric bootstrap method allows to nicely reflect the null hypothesis and to reconstruct the original design. A major drawback, however, is that it fully relies on the possibly misspecified pdf $f_i(y, \theta)$. If the assumed model is not correct, the superior performance of the parametric bootstrap test might collapse (see, e.g., Lee 1994 who studied a procedure for choosing between the parametric and nonparametric bootstrap).

This parametric approach has been applied by Aerts and Claeskens (1999) to pseudo-likelihood models for clustered binary data. In case the model is correctly specified, that is $g_i(y) = f_i(y, \theta)$, but maximizing the pseudolikelihood instead of the true likelihood, they show that this bootstrap approximation can be used as an interesting alternative to the classical asymptotic distribution of estimators and test statistics. In their simulations it is observed that the extremely high significance levels of the chi-squared approximation to the distribution of the
Wald and pseudolikelihood ratio test statistics is nicely corrected by applying the parametric bootstrap resampling scheme.

By resampling the data (nonparametric bootstrap) or residuals (semiparametric bootstrap), no likelihood model has to be specified and therefore this bootstrap approach seems to be preferable in the context of misspecification. A main difficulty however is to construct a resampling scheme that reflects the null hypothesis. Resampling data blindly leads to tests with very low power (see e.g. Hall and Wilson 1991). As a consequence, most bootstrap results focus on the construction of confidence sets (see e.g. Chapter 4 in Shao and Tu 1995).

A nice application of a semiparametric bootstrap test is given in Mammen (1993). For linear models of the form $Y = X\beta + \varepsilon$ and the hypothesis $H_0 : C\beta = h$, he proved the asymptotic correctness of a bootstrap test based on the residual bootstrap. That is, new data are generated as $Y^* = X\hat{\beta}_0 + \varepsilon^*$ with $\hat{\beta}_0$ the restricted least squares estimator under $H_0$ and $\varepsilon^*$ a resample taken with replacement from the residuals (corresponding to the fitted unrestricted model). The residual bootstrap has also been successfully used for generalized linear models, see, e.g., Moulton and Zeger (1989, 1991). For multiparameter (pseudo-)likelihood models, it is not clear how to define residuals that can be used for resampling purposes. There is a need for a method avoiding the introduction of residuals.

Therefore we propose to resample the score and the differentiated score values. Based on a linear approximation, we define a bootstrap replicate of $\hat{\theta}_n$ under $H_0$ as

$$\hat{\theta}_n^* = \hat{\theta}_n^{(0)} - \left( \sum_{i=1}^{p} \sum_{j=1}^{n_i} \psi_{ij}^*(\hat{\theta}_n) \right)^{-1} \sum_{i=1}^{p} \sum_{j=1}^{n_i} \psi_{ij}^*(\hat{\theta}_n)$$

(9)

where, for each $i = 1, \ldots, p$, $(\psi_{ij}(\hat{\theta}_n), \dot{\psi}_{ij}(\hat{\theta}_n))$, $j = 1, \ldots, n_i$ is a sample with replacement from the set $\{(\psi_i(Y_{ij}, \hat{\theta}_n), (\partial/\partial \theta)\dot{\psi}_i(Y_{ij}, \hat{\theta}_n)), j = 1, \ldots, n_i\}$. Note that $\psi_i(Y_{ij}, \hat{\theta}_n)$ is a $r \times 1$ vector and $(\partial/\partial \theta)\dot{\psi}_i(Y_{ij}, \hat{\theta}_n)$ is a $r \times r$ matrix. A similar linearization idea is used in simulation approaches for the bootstrap, as the linear bootstrap (Davison, Hinkley and Schechtman 1986) and the one-step bootstrap (Schucany and Wang 1991). For linear models $Y = X\beta + \varepsilon$, the idea of resampling scores has also been proposed by Hu and Zidek (1995).

The rationale behind definition (9) is as follows. The first term at the right-hand side of (9) reflects the null hypothesis and the second term represents the random fluctuation of the bootstrap replicate $\hat{\theta}_n^*$ around the estimator $\hat{\theta}_n^{(0)}$. The score values are evaluated in the unrestricted estimator $\hat{\theta}_n$, because this term should catch the random mechanism properly, even if the null hypothesis is not true.

The bootstrap Wald and score test statistics based on $\hat{\theta}_n^*$ as defined in (9), coincide and are given by

$$W_n^* = S_n^* = n(\hat{\theta}_n^* - \hat{\theta}_n^{(0)})^T_L \left( A_n^*(\hat{\theta}_n)^{-1} B_n^*(\hat{\theta}_n) A_n^*(\hat{\theta}_n)^{-1} \right)^{-1} L_L (\hat{\theta}_n^* - \hat{\theta}_n^{(0)})_L,$$
where $A_n^*(\theta)$ and $B_n^*(\theta)$ are defined the same way as $A_n(\theta)$ and $B_n(\theta)$, but using the bootstrap scores and derivatives of the scores $(\psi_{ij}^*(\hat{\theta}_n), \psi_{ij}^{**}(\hat{\theta}_n))$, $j = 1, \ldots, n$, instead of $(\psi_i(Y_{ij}, \hat{\theta}_n), (\partial/\partial \theta)\psi_i(Y_{ij}, \hat{\theta}_n))$, $j = 1, \ldots, n$.

The motivation for defining $S_n^*$ equal to $W_n^*$ follows from the classical arguments in proving the asymptotic normality of the score test statistic. It is well known that both test statistics are first order equivalent. A typical way of obtaining the asymptotic distribution of the score test statistic is by substituting $\hat{\theta}_{nL}$ in the Wald statistic by the first $L$ components of the second term in (9). By definition (9) of the one-step linear estimator, this substitution is exact; for estimators in general, this is only approximate.

Definition (9) follows from a linear approximation of the score equations. One might improve on this by including quadratic and higher order terms. A possible approach is suggested by the second and third order efficient approximations as discussed in, e.g., Ghosh (1994). We focus attention on the following second order approximation (simplified to one population),

$$0 = \sum_{j=1}^{n} \psi(Y_{j}, \theta) + \sum_{k=1}^{r} \sum_{j=1}^{n} \frac{\partial}{\partial \theta_k} \psi(Y_{j}, \theta)(\hat{\theta}_{nk} - \theta_k)$$

$$+ \sum_{k=1}^{r} \sum_{\ell=1}^{r} \sum_{j=1}^{n} \frac{\partial^2}{\partial \theta_k \partial \theta_\ell} \psi(Y_{j}, \theta)(\hat{\theta}_{nk} - \theta_k)(\hat{\theta}_{n\ell} - \theta_\ell) + O_P(n^{-1/2}). \quad (10)$$

By calculations similar to those of Ghosh (1994), the expansion (10) suggests the following one-step quadratic estimator

$$\hat{\theta}_n^* = \hat{\theta}_n^{(0)} + U_n^* - \frac{1}{2} \left( \sum_{j=1}^{n} \psi_{ij}^*(\hat{\theta}_n) \right)^{-1} \sum_{k=1}^{r} \sum_{\ell=1}^{r} \sum_{j=1}^{n} \psi_{ij}^{**}(\hat{\theta}_n)_{k,\ell} U_{nk}^* U_{n\ell}^* \quad (11)$$

with

$$U_n^* = - \left( \sum_{i=1}^{p} \sum_{j=1}^{n} \psi_{ij}^*(\hat{\theta}_n) \right)^{-1} \sum_{i=1}^{p} \sum_{j=1}^{n} \psi_{ij}^*(\hat{\theta}_n).$$

This bootstrap estimator is based on the values $(\psi_{ij}^*(\hat{\theta}_n), \psi_{ij}^{**}(\hat{\theta}_n))$, $j = 1, \ldots, n$ taken with replacement from the set

$\{ (\psi(Y_{j}, \hat{\theta}_n), (\partial/\partial \theta)\psi(Y_{j}, \hat{\theta}_n), (\partial^2/\partial \theta \partial \theta^T)\psi(Y_{j}, \hat{\theta}_n)), j = 1, \ldots, n \}.$

It is expected that the last term at the right-hand side of (11) improves the representation of the random variation about the null estimate $\hat{\theta}_n^{(0)}$. This is confirmed in the simulation study.

Both bootstrap procedures (linear and quadratic) lead to consistent estimators for the null distribution of the robust Wald and score test statistics. For the necessary regularity conditions and proofs, we refer to Claeskens (1999).
5 Simulations and the NTP data

5.1 The simulation setting

This section illustrates the finite sample behaviour of the robust chi-squared and the bootstrap Wald and score statistics. Using the models described in Section 3, we examine the performance of the different tests for the no dose effect null hypothesis. For this null hypothesis, the “nearest” member of an incorrectly specified parametric family still satisfies the null hypothesis. The interpretation of this statement should be clear: if there is no effect of the dose in one model, there also will be no dose effect in any other model, and this is independent of the interpretation of the model parameters.

Note that although models of quite different structure are being contrasted, the problem we are looking at really makes a lot of sense. It is exactly what happens in daily practice, since one (almost) never can be sure about the probability model which generated the set of data at hand.

In the simulations we will reconstruct some realistic situations from developmental toxicity studies. A typical toxicological experiment includes one control group and some active dose groups. For the simulations we selected dose levels 0, 0.25, 0.5 and 1. Several parameter settings were investigated, all of these might occur in practical experiments. For each of the selected models we will construct a (robustified) Wald and score statistic for testing the null hypothesis. The simulated levels and size adjusted powers of these tests are compared with those obtained by applying the bootstrap, as explained above. For the no dose effect null hypothesis we might consider two different resampling schemes. The first method is the one as described before, where scores and differentiated scores are resampled for each dose level separately. We denote this by $B_1/D$ for the linear one-step approximation, or by $B_2/D$ when the quadratic approximation is being constructed. For the no dose effect null hypothesis, an alternative valid resampling scheme is to ignore the presence of the dose and to resample from the complete set of scores and differentiated scores. This resampling scheme is denoted by $B_i/A$ ($i = 1, 2$).

An equal number of 15 clusters was assigned to each dose group. First, the number of fetuses $m$ is assumed to be fixed to the value $m = 12$. In a second setting the cluster sizes are random, in that case $m$ is assumed to follow a local linear smoothed version of the relative frequency distribution given in Kupper et al. (1986) (see Table 1 in Molenberghs et al., 1998). Ideally, when several different litter sizes are observed, resampling should be done within each group of litters with identical dosage and identical litter size (being a separate population). Since litter sizes typically vary from 1 to about 20, this would lead to very small sample sizes for several of these populations and resampling would not be very effective. In our simulations with random litter sizes and also for the NTP data, we resampled scores and differentiated scores from all litters.
in the same dose group (having different sizes) or from the complete set (as explained before). In this way a bootstrap estimate $\hat{\theta}_n^*$ is not based on contributions of litters with exactly the same size distribution as in the original sample, but asymptotically it does reflect the (unknown) littersize distribution.

For each setting 500 datasets were generated using the build-in GAUSS routine RNDU. On each dataset the $P$-values of the robust Wald ($W_n$) and the robust score ($S_n$) statistics were computed based on their limiting $\chi^2$ distribution and on the simulated bootstrap distribution, using 1000 bootstrap resamples.

### 5.2 Simulation results

As a first case, data were generated from the conditional MR model with intercept $\theta_{10} = -1.5, -1.0, -0.5$, dose effect $\theta_{11} = 0, 0.5$ and association parameter $\theta_{20} = 0.1, 0.2$. The beta-binomial model was used to fit the data and to test the no dose null hypothesis at the 5% level of significance. Table 1 shows simulated type I errors (as percentages) and Table 2 shows the simulated power of the tests.

#### Tables 1 and 2 About Here

In this setting we only considered the linear approximation bootstrap tests. In principle, one could define a quadratic test too, but the mathematical calculations are getting rather cumbersome for the beta-binomial model.

First, we observe that the robust Wald and score $\chi^2$ tests seem to behave very similar with only slightly inflated type I errors. There is not much room for improvement by the bootstrap but the linear one-step bootstrap pulls the inflated type I errors of both robust $\chi^2$-tests down. If the scores are resampled from the complete set of scores (ignoring the dose levels), the $B_1/A$ bootstrap tests tend to be somewhat conservative. For the results on power characteristics to be comparable for the different tests, we show the size-adjusted rejection probabilities in Table 2. These results indicate that the loss in power for the bootstrap methods is almost negligible, especially for the $B_1/A$ test. In summary, for this setting, only a small correction in the level of $\chi^2$ tests is necessary, which is achieved by the one-step bootstrap.

In a second case, data were generated using a beta-binomial model with parameters $\theta_{10} = -4, -3.5, -3.0, -2.5, \theta_{11} = 0, 1.0 , \theta_{20} = .2 , .3$. Fitting and testing was based on the pseudolikelihood model (5). Now there are two sources of misspecification: the assumed probability model is wrong and the pseudolikelihood technique has been used, instead of full likelihood estimation.

#### Table 3 About Here
Type I errors and rejection probabilities were simulated for the $\chi^2$ Wald and score tests, and the linear and quadratic bootstrap tests for both resampling schemes ($B_i/D, B_i/A$, $i = 1, 2$). As in the first setting, the type I errors are larger than the nominal level but now it is more pronounced, especially for smaller values of $\theta_{10}$ (determining the baseline malformation probability for zero dose) and for higher intra-litter association represented by $\theta_{20}$. This latter situation corresponds to the case of the least expected number of events (malformations) and the least sample information (similar behaviour of subjects within the same litter). Now, robust Wald and score $\chi^2$ tests behave differently. In fact, the robust score $\chi^2$ test is doing very well. The one-step linear bootstrap test is hardly any better than its $\chi^2$ counterpart but the quadratic bootstrap Wald tests (both $B_2/D$ and $B_2/A$) seem to nicely correct the level downwards. Also for the single setting in which the score test has a small problem, the bootstrap works. Note that the robust score statistic for the quadratic bootstrap test coincides, by definition of a score test, with the statistic of the linear bootstrap test. For this reason they are not shown in the tables.

As a comparison, type I errors were also considered for the extremely computer intensive, fully iterative bootstrap Wald test $B_{it}/D$. To obtain the simulation results of this test, we resampled for each of the 500 sets of original data, 1000 times, per dose level (with replacement, on cluster level, sample size 15). The maximum likelihood estimates $\hat{\theta}_{10}$ of each of these 1000 resampled data sets were computed fully iteratively (requiring about 20 iterations) and used to get, for each of the original data sets, 1000 replicates of the Wald test statistic, now defined as $W_n^* = n(\hat{\theta}_{10} - \hat{\theta}_{n})^T(L_n^*(\hat{\theta}_{n})^{-1}B_n^*(\hat{\theta}_{n})A_n^*(\hat{\theta}_{n})^{-1})^{-1}L(\hat{\theta}^* - \hat{\theta}_{n})_n$. This definition follows the guidelines of Hall and Wilson (1991). A well-known drawback of this bootstrap method is the fact that it nowhere reflects the null hypothesis. Related to that, one might expect a substantial loss in power. From Table 3 it follows that, for this application, the method has a really poor performance. Moreover, it is computationally extremely time consuming.

Table 4 gives some indication of loss of power for the one-step bootstrap tests. This loss is more pronounced for the linear bootstrap. The quadratic bootstrap tests seem to lift up the power close to that of the $\chi^2$ tests. In conclusion, for the robust Wald test, we recommend the quadratic one-step bootstrap approach. Its level is close to the nominal level, and it has good power characteristics. There is no substantial difference between the two resampling schemes $B_i/D$ and $B_i/A$. The robust score $\chi^2$ test is behaving very well (also having larger power) and the bootstrap can only confirm this good performance.

Table 4 About Here

For the more realistic situation that the cluster sizes vary, the picture remains more or less the same. This is illustrated in Tables 5 and 6 where the same settings as in Tables 3 and 4...
were used. The level of the Wald test, using chi-squared critical points, is often too high and
only the quadratic bootstrap test is able to correct this sufficiently. Also for random litter sizes,
the robustified score test is the one to be preferred. We also examined this last setting using
the MR instead of the pseudolikelihood model. Previous conclusions are confirmed, so results
are not shown here (for details, see Claeskens 1999).

Table 5 and 6 About Here

5.3 The NTP data

Lindstrom et al. (1990) investigated the effect in mice of the chemical theophylline. In this
experiment, there were 4 dose levels: 0, 0.375, 0.75, and 1.0. For these doses there were,
respectively, 25, 25, 29 and 17 litters and the sizes of the litters were ranging from 2 to 16.
Malformations were classified as being external, visceral and skeletal. Also a collapsed outcome
is considered, which indicates whether a fetus has at least one malformation.

In each of the three models, the beta-binomial (BB) model (6), the conditional model (3) of
Molenberghs and Ryan (MR) and the pseudolikelihood (PL) model (5), we used parameterisation
(4). We are interested in testing the no dose effect hypothesis on the main effect parameter
\( H_0 : \theta_{11} = 0 \).

Table 7 About Here

The results are shown in Table 7. The table shows the \( P \)-values (as \% ) of the different test
statistics discussed before. We also included the results from a GEE2 estimation method based
on the first four order moments of the Bahadur model. The Bahadur model for clustered binary
data has the same first- and second-order moments as the beta-binomial model. For more details,
we refer to Bahadur (1961) and Kupper and Haseman (1978) for the Bahadur model and to
Zhao and Prentice (1990) for the GEE2 estimation method.

For external malformations, the results of the different tests are almost the same for the beta-
binomial model and close to each other for the other models. There seems to be no significant
effect of theophylline on the external malformation probability.

For visceral and skeletal malformations, there is a striking discrepancy between the Wald and
the score test. The significance of the Wald test should be interpreted with care (see also the
inflated type I errors in Tables 1, 3 and 5). The quadratic bootstrap seems to correct the Wald
test in the direction of the score test. Compared with the chi-squared tests, the bootstrap score
test has higher \( P \)-values for visceral malformation and lower \( P \)-values for skeletal malformation.
For the beta-binomial model there were convergence problems when fitting the null model. For
this reason, the score statistics could not be obtained. Also for the GEE2 model some problems
arose, but these could be avoided by using a Moore-Penrose generalized inverse of the matrix
For visceral malformations, the null hypothesis can not be rejected. A conclusion is less clear for skeletal malformations. Except for the quadratic bootstrap for the MR model, all Wald tests indicate a significant dose effect, while all score tests indicate no effect (although the bootstrap score PL test is getting close to 5%). Since the score tests showed to have a good behaviour in our simulation study, we might believe the results of these test, though further investigation might be necessary to come to a conclusion. Finally, also note that, for all three types of malformation, the different Wald tests lead to highly variable $P$-values whereas the score tests are much more stable.

The collapsed version seems to indicate that theophylline might have an effect on the development of fetuses. Here, all score $P$-values are between 0.0220 and 0.0754 and the Wald $P$-values are between 0.0150 and 0.0820. This indicates that a separate analysis of each type of malformation can lead to misleading conclusions and that for this type of problems one has to consider all types jointly as a multivariate respons or at least, as we did here, a collapsed malformation indicator.

6 Discussion

In this paper we examine the behaviour of the robust Wald and score test for clustered binary data, based on (pseudo-)likelihood estimation and allowing for misspecification. A new semi-parametric bootstrap test based on resampling scores and their derivatives is proposed and is contrasted with the classical chi-squared robust tests. Asymptotically the bootstrap method is consistent and finite sample simulations show substantial improvements in size, especially for the quadratic bootstrap Wald test. This quadratic one-step method is easily obtained for exponential family based distributions, but calculations can be more difficult for other distributions. Since our simulations showed the poor behaviour of the chi-squared Wald test and suggested good results for the chi-squared score test, this latter test might be a good choice to use for statistical analysis. Unfortunately, most of the existing statistical software packages don’t provide this score test automatically.

Although the technique is introduced in the context of (pseudo-)likelihood, it can also be applied for estimating equations in general, as is being illustrated for the data on theophylline. Currently we are investigating other domains of application of this bootstrap method, such as the construction of confidence intervals.

Acknowledgements

This project was partly supported by the NATO Collaborative Research Grant 950648. The research of Gerda Claeskens is supported by the Fund for Scientific Research - Flanders (Belgium) (FWO).
References


Table 1
Simulated type I errors (as %), significance level 0.05.
Data are generated with the MR model and fitted using the beta-binomial model, clustersize=12. $H_0: \theta_{11} = 0$.

<table>
<thead>
<tr>
<th>$\theta_{10}$</th>
<th>$\theta_{20} = 0.1$</th>
<th>$\theta_{20} = 0.2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\chi^2$</td>
<td>$B_1/D$</td>
</tr>
<tr>
<td>-1.5</td>
<td>$W_n$</td>
<td>7.20*</td>
</tr>
<tr>
<td></td>
<td>$S_n$</td>
<td>6.40</td>
</tr>
<tr>
<td></td>
<td>$W_n$</td>
<td>7.21*</td>
</tr>
<tr>
<td></td>
<td>$S_n$</td>
<td>8.02*</td>
</tr>
<tr>
<td>-0.5</td>
<td>$W_n$</td>
<td>7.80*</td>
</tr>
<tr>
<td></td>
<td>$S_n$</td>
<td>7.80*</td>
</tr>
</tbody>
</table>

* denotes the proportion of significant tests (at 5%) which differs significantly from 5%

Table 2
Simulated power (as %), significance level 0.05.
Data are generated with the MR model and fitted using the beta-binomial model, clustersize=12. $H_0: \theta_{11} = 0$.

<table>
<thead>
<tr>
<th>$\theta_{11} = 0.5$</th>
<th>$\theta_{20} = 0.1$</th>
<th>$\theta_{20} = 0.2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\chi^2$</td>
<td>$B_1/D$</td>
</tr>
<tr>
<td>-1.5</td>
<td>$W_n$</td>
<td>39.7</td>
</tr>
<tr>
<td></td>
<td>$S_n$</td>
<td>36.7</td>
</tr>
<tr>
<td>-1.0</td>
<td>$W_n$</td>
<td>62.3</td>
</tr>
<tr>
<td></td>
<td>$S_n$</td>
<td>63.5</td>
</tr>
<tr>
<td>-0.5</td>
<td>$W_n$</td>
<td>87.2</td>
</tr>
<tr>
<td></td>
<td>$S_n$</td>
<td>83.7</td>
</tr>
</tbody>
</table>
Table 3
Simulated type I errors (as %), significance level 0.05.
Data are generated with the beta-binomial model and fitted
using the pseudolikelihood model, clustersize=12. $H_0 : \theta_{11} = 0.$

<table>
<thead>
<tr>
<th>$\theta_{10}$</th>
<th>$\chi^2$</th>
<th>$B_1/D$</th>
<th>$B_2/D$</th>
<th>$B_{it}/D$</th>
<th>$B_1/A$</th>
<th>$B_2/A$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_{20} = 0.2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-4.0</td>
<td>$W_n$</td>
<td>10.55*</td>
<td>10.76*</td>
<td>6.96</td>
<td>10.76*</td>
<td>9.28*</td>
</tr>
<tr>
<td>$S_n$</td>
<td>6.12</td>
<td>6.75</td>
<td>—</td>
<td>—</td>
<td>5.70</td>
<td>—</td>
</tr>
<tr>
<td>-3.5</td>
<td>$W_n$</td>
<td>5.65</td>
<td>6.45</td>
<td>4.23</td>
<td>5.85</td>
<td>5.44</td>
</tr>
<tr>
<td>$S_n$</td>
<td>4.64</td>
<td>3.83</td>
<td>—</td>
<td>—</td>
<td>3.02</td>
<td>—</td>
</tr>
<tr>
<td>-3.0</td>
<td>$W_n$</td>
<td>6.40</td>
<td>5.80</td>
<td>5.20</td>
<td>7.00*</td>
<td>5.60</td>
</tr>
<tr>
<td>$S_n$</td>
<td>6.80</td>
<td>5.40</td>
<td>—</td>
<td>—</td>
<td>4.60</td>
<td>—</td>
</tr>
<tr>
<td>-2.5</td>
<td>$W_n$</td>
<td>7.80*</td>
<td>6.60</td>
<td>5.80</td>
<td>8.20*</td>
<td>6.00</td>
</tr>
<tr>
<td>$S_n$</td>
<td>7.40*</td>
<td>5.60</td>
<td>—</td>
<td>—</td>
<td>5.20</td>
<td>—</td>
</tr>
</tbody>
</table>

| $\theta_{20} = 0.3$ |
|----------------|---------|---------|---------|------------|---------|---------|
| -4.0 | $W_n$ | 10.71* | 11.65* | 7.07* | 11.35* | 10.71* | 5.78 |
| $S_n$ | 6.21 | 6.21 | — | — | 4.71 | — |
| -3.5 | $W_n$ | 7.06* | 7.66* | 5.04 | 7.46* | 6.25 | 4.64 |
| $S_n$ | 6.86 | 5.04 | — | — | 4.44 | — |
| -3.0 | $W_n$ | 7.80* | 7.20* | 5.20 | 8.00* | 6.60 | 5.20 |
| $S_n$ | 6.60 | 5.60 | — | — | 5.00 | — |
| -2.5 | $W_n$ | 6.40 | 6.00 | 5.80 | 6.40 | 5.80 | 5.40 |
| $S_n$ | 6.80 | 5.20 | — | — | 4.40 | — |

* denotes the proportion of significant tests (at 5%) which differs significantly from 5%.
Table 4
Simulated power (as %), significance level 0.05. Data are generated with the beta-binomial model and fitted using the pseudolikelihood model, clustersize=12, $\theta_{11} = 1; H_0 : \theta_{11} = 0$.

<table>
<thead>
<tr>
<th>$\theta_{10}$</th>
<th>$\chi^2$</th>
<th>$B_1/D$</th>
<th>$B_2/D$</th>
<th>$B_1/A$</th>
<th>$B_2/A$</th>
<th>$\theta_{20} = 0.2$</th>
<th>$\chi^2$</th>
<th>$B_1/D$</th>
<th>$B_2/D$</th>
<th>$B_1/A$</th>
<th>$B_2/A$</th>
<th>$\theta_{20} = 0.3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-4.0 W_n</td>
<td>15.2</td>
<td>8.4</td>
<td>8.9</td>
<td>8.8</td>
<td>9.8</td>
<td>11.6</td>
<td>11.3</td>
<td>7.0</td>
<td>12.7</td>
<td></td>
<td></td>
<td>15.9</td>
</tr>
<tr>
<td>S_n</td>
<td>19.0</td>
<td>13.2</td>
<td>—</td>
<td>14.5</td>
<td>—</td>
<td>15.9</td>
<td>10.1</td>
<td>—</td>
<td>22.0</td>
<td></td>
<td></td>
<td>25.4</td>
</tr>
<tr>
<td>-3.5 W_n</td>
<td>30.2</td>
<td>24.9</td>
<td>30.2</td>
<td>26.2</td>
<td>27.5</td>
<td>25.2</td>
<td>17.4</td>
<td>23.4</td>
<td>21.0</td>
<td>23.5</td>
<td></td>
<td>25.4</td>
</tr>
<tr>
<td>S_n</td>
<td>33.8</td>
<td>30.6</td>
<td>—</td>
<td>30.8</td>
<td>—</td>
<td>25.4</td>
<td>25.0</td>
<td>—</td>
<td>34.6</td>
<td></td>
<td></td>
<td>37.1</td>
</tr>
<tr>
<td>-3.0 W_n</td>
<td>41.0</td>
<td>34.7</td>
<td>35.2</td>
<td>33.8</td>
<td>36.6</td>
<td>29.1</td>
<td>23.6</td>
<td>29.5</td>
<td>24.9</td>
<td>27.7</td>
<td></td>
<td>37.1</td>
</tr>
<tr>
<td>S_n</td>
<td>43.7</td>
<td>39.8</td>
<td>—</td>
<td>42.8</td>
<td>—</td>
<td>37.1</td>
<td>31.3</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
<td>37.1</td>
</tr>
<tr>
<td>-2.5 W_n</td>
<td>51.5</td>
<td>44.6</td>
<td>48.1</td>
<td>45.5</td>
<td>47.6</td>
<td>45.2</td>
<td>38.8</td>
<td>40.7</td>
<td>38.0</td>
<td>39.7</td>
<td></td>
<td>49.4</td>
</tr>
<tr>
<td>S_n</td>
<td>57.3</td>
<td>54.3</td>
<td>—</td>
<td>53.9</td>
<td>—</td>
<td>49.4</td>
<td>48.9</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
<td>50.0</td>
</tr>
</tbody>
</table>

Table 5
Simulated type I errors (as %), significance level 0.05. Data are generated with the beta-binomial model and fitted using the pseudolikelihood model. $H_0 : \theta_{11} = 0$. Random clustersizes.

<table>
<thead>
<tr>
<th>$\theta_{10}$</th>
<th>$\chi^2$</th>
<th>$B_1/D$</th>
<th>$B_2/D$</th>
<th>$B_1/A$</th>
<th>$B_2/A$</th>
<th>$\theta_{20} = 0.2$</th>
<th>$\chi^2$</th>
<th>$B_1/D$</th>
<th>$B_2/D$</th>
<th>$B_1/A$</th>
<th>$B_2/A$</th>
<th>$\theta_{20} = 0.3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-4.0 W_n</td>
<td>9.84*</td>
<td>9.04*</td>
<td>6.02</td>
<td>9.64*</td>
<td>5.62</td>
<td>11.72*</td>
<td>10.71*</td>
<td>4.85</td>
<td>9.70*</td>
<td>4.65</td>
<td></td>
<td>3.84</td>
</tr>
<tr>
<td>S_n</td>
<td>5.62</td>
<td>4.62</td>
<td>—</td>
<td>3.41</td>
<td>—</td>
<td>3.84</td>
<td>4.04</td>
<td>—</td>
<td>4.80</td>
<td></td>
<td></td>
<td>6.40</td>
</tr>
<tr>
<td>-3.5 W_n</td>
<td>8.40*</td>
<td>7.80*</td>
<td>4.80</td>
<td>6.60</td>
<td>4.40</td>
<td>9.00*</td>
<td>8.60*</td>
<td>4.80</td>
<td>7.80*</td>
<td>4.80</td>
<td></td>
<td>6.40</td>
</tr>
<tr>
<td>S_n</td>
<td>5.80</td>
<td>5.80</td>
<td>—</td>
<td>4.20</td>
<td>—</td>
<td>6.40</td>
<td>5.80</td>
<td>—</td>
<td>4.80</td>
<td></td>
<td></td>
<td>6.40</td>
</tr>
<tr>
<td>-3.0 W_n</td>
<td>8.40*</td>
<td>7.40*</td>
<td>5.60</td>
<td>6.80</td>
<td>5.40</td>
<td>8.40*</td>
<td>6.80</td>
<td>5.20</td>
<td>5.80</td>
<td>4.20</td>
<td></td>
<td>6.40</td>
</tr>
<tr>
<td>S_n</td>
<td>6.40</td>
<td>5.80</td>
<td>—</td>
<td>5.00</td>
<td>—</td>
<td>6.40</td>
<td>4.60</td>
<td>—</td>
<td>3.80</td>
<td></td>
<td></td>
<td>6.40</td>
</tr>
<tr>
<td>-2.5 W_n</td>
<td>6.00</td>
<td>5.00</td>
<td>4.40</td>
<td>5.00</td>
<td>4.20</td>
<td>6.80</td>
<td>5.60</td>
<td>4.00</td>
<td>5.40</td>
<td>3.60</td>
<td></td>
<td>6.80</td>
</tr>
<tr>
<td>S_n</td>
<td>6.00</td>
<td>4.60</td>
<td>—</td>
<td>4.00</td>
<td>—</td>
<td>6.00</td>
<td>4.40</td>
<td>—</td>
<td>3.80</td>
<td></td>
<td></td>
<td>6.00</td>
</tr>
</tbody>
</table>

* denotes the proportion of significant tests (at 5%) which differs significantly from 5%
**Table 6**

Simulated power (as %), significance level 0.05. Data are generated with the beta-binomial model and fitted using the pseudolikelihood model.

$\theta_{11} = 1$; $H_0 : \theta_{11} = 0$. Random clustersizes.

<table>
<thead>
<tr>
<th>$\theta_{10}$</th>
<th>$\theta_{20} = 0.2$</th>
<th>$\theta_{20} = 0.3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$W_n$</td>
<td>$\chi^2$</td>
<td>$B_1/D$</td>
</tr>
<tr>
<td>-4.0</td>
<td>14.5</td>
<td>9.8</td>
</tr>
<tr>
<td>$S_n$</td>
<td>18.4</td>
<td>15.7</td>
</tr>
<tr>
<td>-3.5</td>
<td>26.0</td>
<td>24.4</td>
</tr>
<tr>
<td>$S_n$</td>
<td>27.5</td>
<td>23.4</td>
</tr>
<tr>
<td>-3.0</td>
<td>34.2</td>
<td>29.4</td>
</tr>
<tr>
<td>$S_n$</td>
<td>36.2</td>
<td>34.6</td>
</tr>
<tr>
<td>-2.5</td>
<td>51.2</td>
<td>51.1</td>
</tr>
<tr>
<td>$S_n$</td>
<td>54.4</td>
<td>52.4</td>
</tr>
</tbody>
</table>
### Table 7

Analysis of the NTP data on theophylline with $H_0: \theta_{11} = 0$. $P$-values are shown as %.

<table>
<thead>
<tr>
<th></th>
<th>External</th>
<th>Visceral</th>
<th>Skeletal</th>
<th>Collapsed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$W_n$</td>
<td>$S_n$</td>
<td>$W_n$</td>
<td>$S_n$</td>
</tr>
<tr>
<td>BB</td>
<td>$\chi^2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>22.85</td>
<td>23.54</td>
<td>4.50*</td>
<td>—</td>
</tr>
<tr>
<td>$B_1/A$</td>
<td>22.78</td>
<td>23.25</td>
<td>0.50*</td>
<td>—</td>
</tr>
<tr>
<td>$B_1/D$</td>
<td>23.98</td>
<td>24.66</td>
<td>0.90*</td>
<td>—</td>
</tr>
<tr>
<td>GEE2</td>
<td>$\chi^2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>14.60</td>
<td>13.90</td>
<td>3.61*</td>
<td>17.27*</td>
</tr>
<tr>
<td>$B_1/A$</td>
<td>15.70</td>
<td>14.40</td>
<td>1.40*</td>
<td>18.80*</td>
</tr>
<tr>
<td>$B_1/D$</td>
<td>14.20</td>
<td>13.30</td>
<td>1.00*</td>
<td>20.20*</td>
</tr>
<tr>
<td>MR</td>
<td>$\chi^2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>9.76</td>
<td>16.98</td>
<td>3.55*</td>
<td>18.26</td>
</tr>
<tr>
<td>$B_1/A$</td>
<td>9.30</td>
<td>18.80</td>
<td>0.60*</td>
<td>21.60</td>
</tr>
<tr>
<td>$B_1/D$</td>
<td>7.80</td>
<td>16.30</td>
<td>0.50*</td>
<td>19.70</td>
</tr>
<tr>
<td>$B_2/A$</td>
<td>8.70</td>
<td>—</td>
<td>5.20</td>
<td>—</td>
</tr>
<tr>
<td>$B_2/D$</td>
<td>7.70</td>
<td>—</td>
<td>3.40*</td>
<td>—</td>
</tr>
<tr>
<td>PL</td>
<td>$\chi^2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>12.26</td>
<td>17.06</td>
<td>3.47*</td>
<td>18.56</td>
</tr>
<tr>
<td>$B_1/A$</td>
<td>12.30</td>
<td>19.10</td>
<td>0.50*</td>
<td>21.70</td>
</tr>
<tr>
<td>$B_1/D$</td>
<td>11.10</td>
<td>16.10</td>
<td>0.50*</td>
<td>19.70</td>
</tr>
<tr>
<td>$B_2/A$</td>
<td>17.60</td>
<td>—</td>
<td>10.30</td>
<td>—</td>
</tr>
<tr>
<td>$B_2/D$</td>
<td>13.90</td>
<td>—</td>
<td>8.60</td>
<td>—</td>
</tr>
</tbody>
</table>

$A*$ denotes rejection at the 5% level and a $\circ$ indicates that a Moore-Penrose generalized inverse is used to obtain the results.